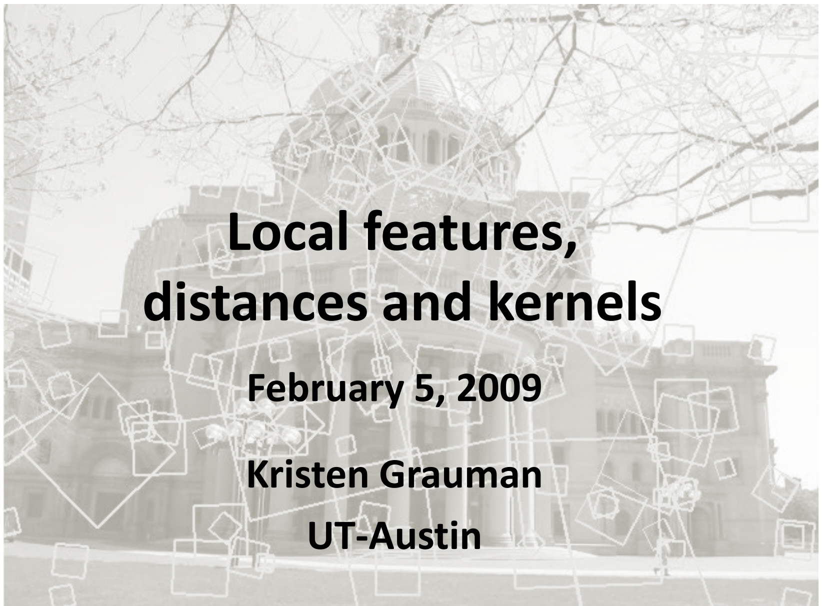


Local features, distances and kernels

February 5, 2009

Kristen Grauman
UT-Austin

Plan for today

- Lecture : local features and matching
- Papers:
  - Video Google [Sivic & Zisserman]
  - Pyramid match [Grauman & Darrell]
  - Learning local distance functions [Frome et al.]
- Demo:
  - Feature sampling strategies for categorization

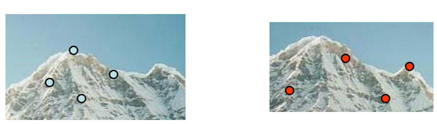
Local features: motivation

- Last week: appearance-based features assuming window under consideration
  - Is fairly aligned across examples
  - Has similar total structure, same components present

- This week: local representations to offer robustness to occlusion, clutter, viewpoint changes,…
  - How to describe
  - How to compare

Local invariant features

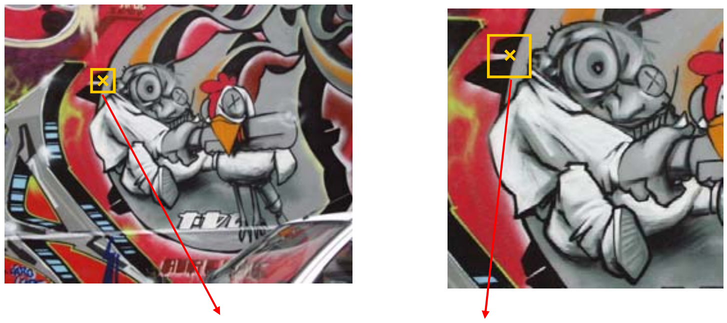
- Problem 1:
  - Detect the *same* point *independently* in both images



no chance to match!

We need a repeatable detector

Automatic Scale Selection



$$f(I_{i_1 \ldots i_m}(x,\sigma)) \;=\; f(I_{i_1 \ldots i_m}(x',\sigma'))$$

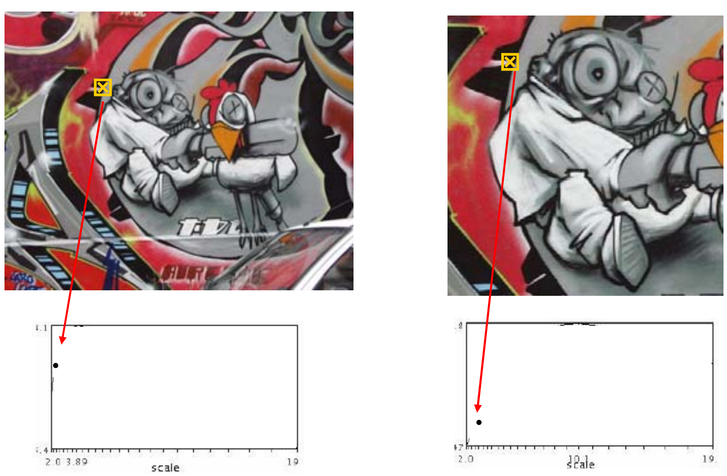
Same operator responses if the patch contains the same image up to scale factor

How to find corresponding patch sizes?

Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Automatic Scale Selection

- **Function responses for increasing scale (scale signature)**

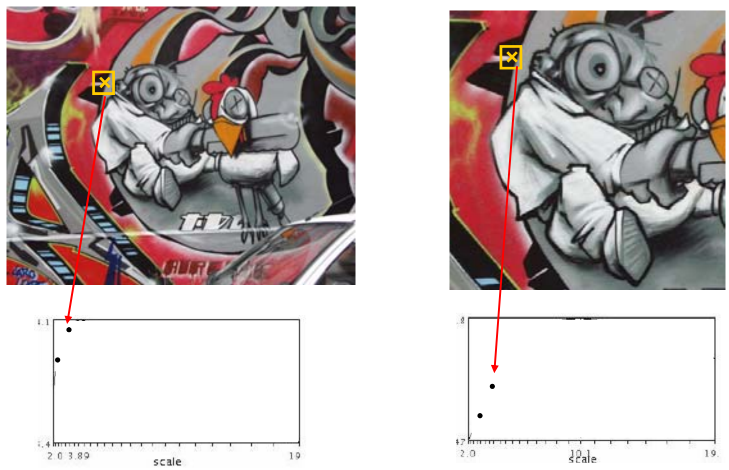


$f(I_{i_1 \ldots i_m}(x,\sigma))$ $\qquad$ $f(I_{i_1 \ldots i_m}(x',\sigma))$

Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Automatic Scale Selection

- **Function responses for increasing scale (scale signature)**

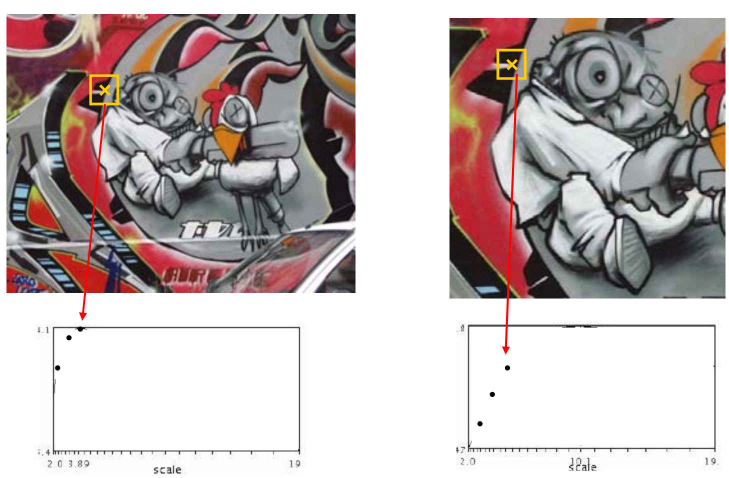


$f(I_{i_1 \ldots i_m}(x,\sigma))$ $\quad$ $f(I_{i_1 \ldots i_m}(x',\sigma))$

Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Automatic Scale Selection

- **Function responses for increasing scale (scale signature)**

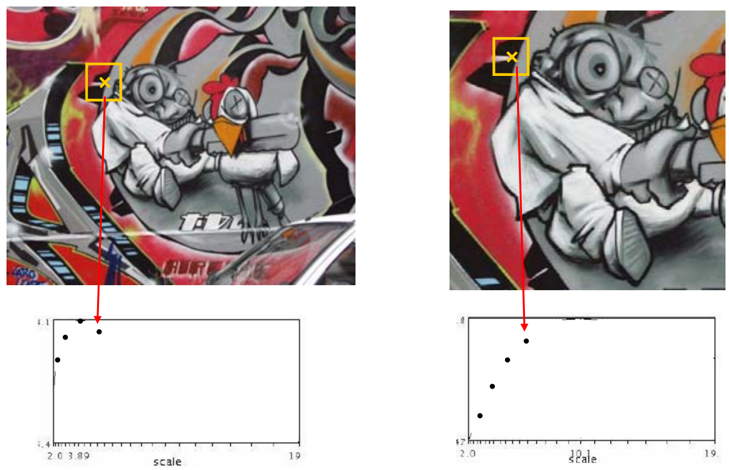


$f(I_{i_1 \ldots i_m}(x,\sigma))$ $\quad$ $f(I_{i_1 \ldots i_m}(x',\sigma))$

Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Automatic Scale Selection

- **Function responses for increasing scale (scale signature)**

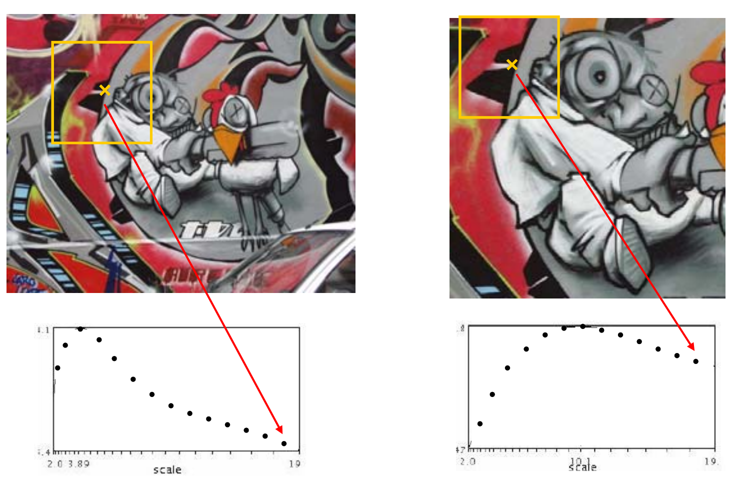


$f(I_{i_1 \ldots i_m}(x,\sigma))$ $f(I_{i_1 \ldots i_m}(x',\sigma))$

Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Automatic Scale Selection

- **Function responses for increasing scale (scale signature)**

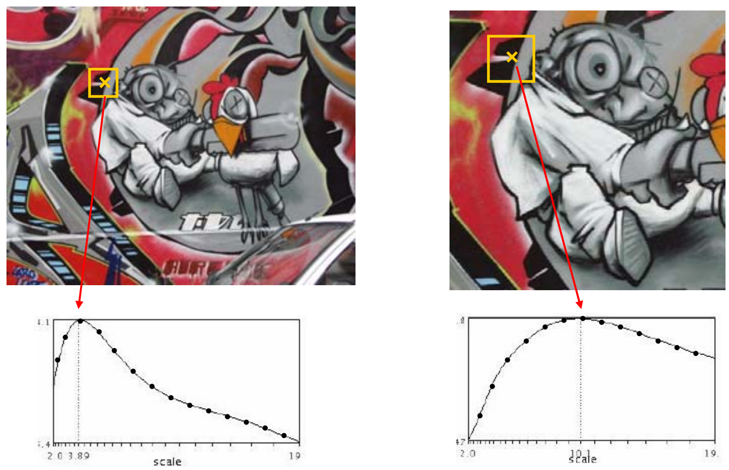


$f(I_{i_1 \ldots i_m}(x,\sigma))$ $f(I_{i_1 \ldots i_m}(x',\sigma))$

Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Automatic Scale Selection

- **Function responses for increasing scale (scale signature)**

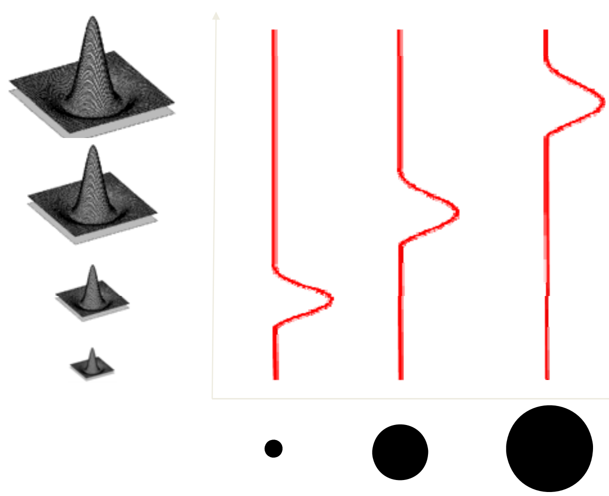


$f(I_{i_1 \dots i_m}(x,\sigma))$ $f(I_{i_1 \dots i_m}(x',\sigma'))$

Slide credit K. Grauman, B. Leibe AAAI08 Short Course

What Is A Useful Signature Function?

- Laplacian-of-Gaussian = "blob" detector



Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Scale-space blob detector: Example

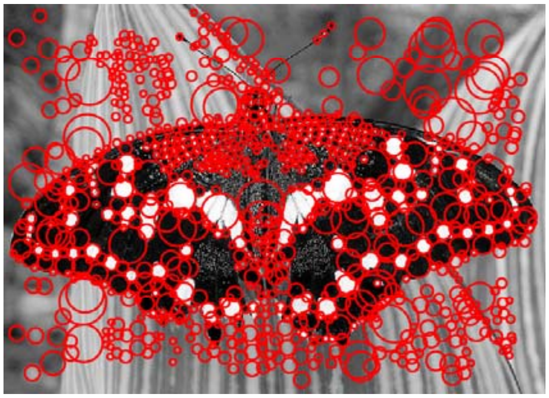
Source: Lana Lazebnik

Scale-space blob detector: Example

sigma = 11.9912

Source: Lana Lazebnik

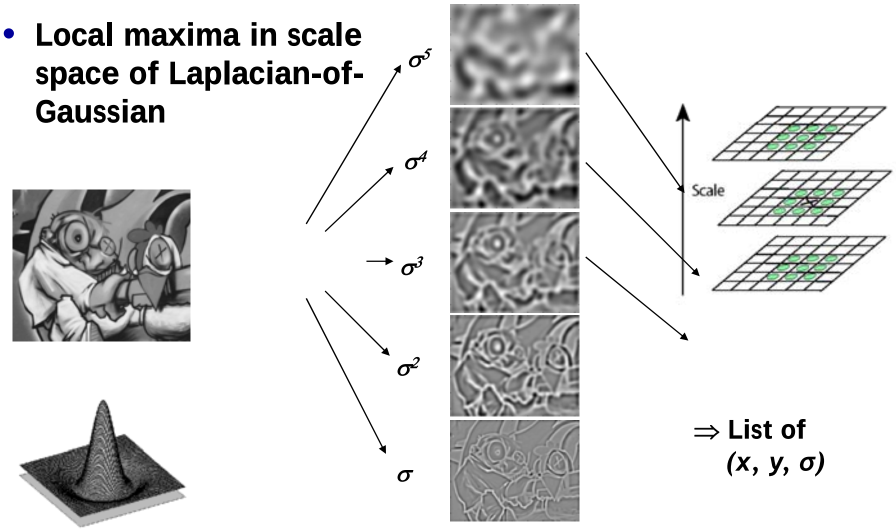
Scale-space blob detector: Example



Source: Lana Lazebnik

Laplacian-of-Gaussian (LoG) for scale invariant detection

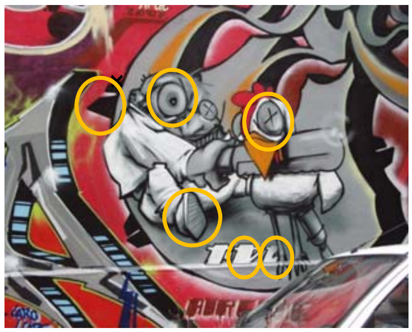
- **Local maxima in scale space of Laplacian-of-Gaussian**



⇒ List of ***(x, y, σ)***

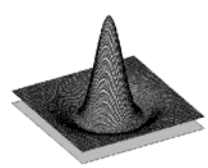
Slide credit K. Grauman, B. Leibe AAAI08 Short Course

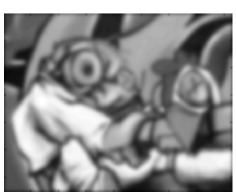
Laplacian of Gaussian: scale invariant detection



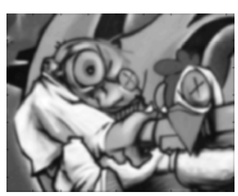
Difference-of-Gaussian (DoG)

- Difference of Gaussians gives an efficient approximation of the Laplacian-of-Gaussian





-



=

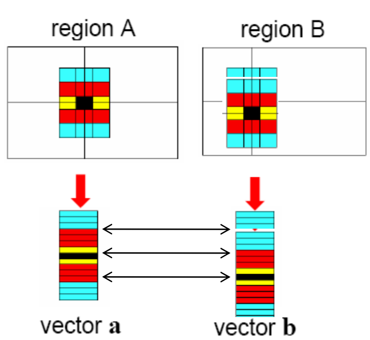
Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Local invariant features

- Problem 2:
  - For each point correctly recognize the corresponding one

We need a reliable and distinctive descriptor

Raw patches as local descriptors



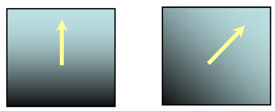
The simplest way to describe the neighborhood around an interest point is to write down the list of intensities to form a feature vector.

But this is very sensitive to even small shifts, rotations.

Rotation invariant descriptors

- Find local orientation

  Dominant direction of gradient for the image patch



- Rotate patch according to this angle

  This puts the patches into a canonical orientation.

What about illumination and translation?

SIFT Descriptor [Lowe 2004]

- Use histograms to bin pixels within sub-patches according to their orientation.
- 4x4x8 = 128 dimensional feature vector

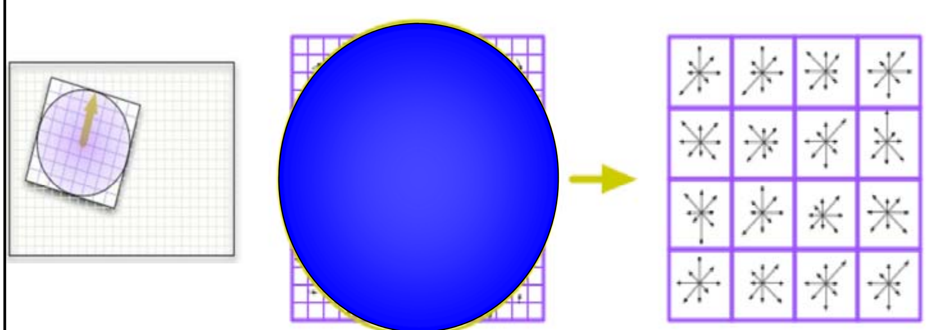


Image from: Jonas Hurrelmann

Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

SIFT Descriptor [Lowe 2004]

Extraordinarily robust matching technique
- Can handle changes in viewpoint
  - Up to about 60 degree out of plane rotation
- Can handle significant changes in illumination
  - Sometimes even day vs. night (below)
- Fast and efficient—can run in real time
- Lots of code available
  - http://people.csail.mit.edu/albert/ladypack/wiki/index.php/Known_implementations_of_SIFT

Slide credit: Steve Seitz

Local invariant features: basic flow

1) Detect interest points
2) Extract descriptors

Descriptors map each region in image to a (typically high-dimensional) feature vector.

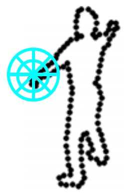
$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_d \end{pmatrix}$$

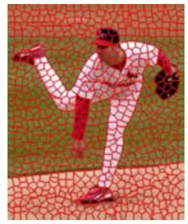
$$\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_d \end{pmatrix}$$

Local representations

Many options for detection & description…


SIFT [Lowe 99]

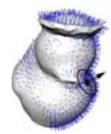

Shape context [Belongie 02]

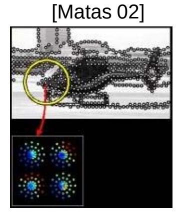

Superpixels [Ren et al.]


Maximally Stable Extremal Regions [Matas 02]


Salient regions [Kadir 01]


Harris-Affine [Mikolajczyk 04]


Spin images [Johnson 99]


Geometric Blur [Berg 05]

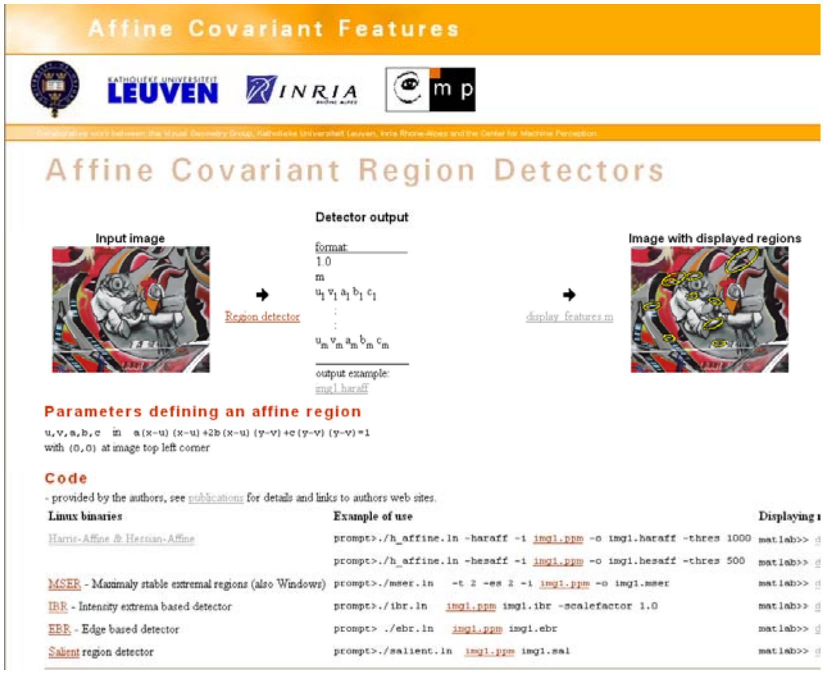
25

Visual Object Recognition Tutorial

You Can Try It At Home…

- **For most local feature detectors, executables are available online:**
- **http://robots.ox.ac.uk/~vgg/research/affine**
- **http://www.cs.ubc.ca/~lowe/keypoints/**
- **http://www.vision.ee.ethz.ch/~surf**

K. Grauman, B. Leibe



http://www.robots.ox.ac.uk/~vgg/research/affine/detectors.html#binaries

Applications of local invariant features & matching

- Wide baseline stereo
- Motion tracking
- Panoramas
- Mobile robot navigation
- 3D reconstruction
- Recognition
  - Specific objects
  - Textures
  - Categories
- …

Wide baseline stereo

[Image from T. Tuytelaars ECCV 2006 tutorial]

Panorama stitching

(a) Matier data set (7 images)

(b) Matier final stitch

Brown, Szeliski, and Winder, 2005

Automatic mosaicing

http://www.cs.ubc.ca/~mbrown/autostitch/autostitch.html

Recognition of specific objects, scenes

Schmid and Mohr 1997

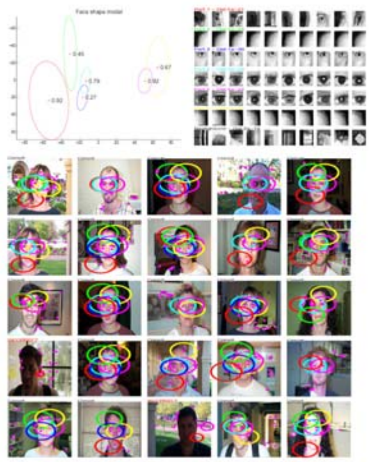
Sivic and Zisserman, 2003

Rothganger et al. 2003

Lowe 2002

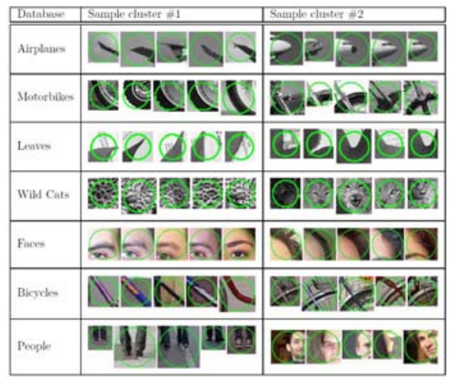
Recognition of categories

Constellation model



Weber et al. (2000)
Fergus et al. (2003)

Bags of words



Csurka et al. (2004)
Dorko & Schmid (2005)
Sivic et al. (2005)
Lazebnik et al. (2006), …

[Slide from Lana Lazebnik, Sicily 2006]

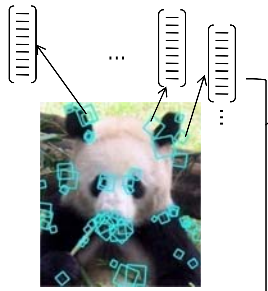
Value of local features

- Critical to find distinctive and repeatable local regions for multi-view matching
- Complexity reduction via selection of distinctive points
- Describe images, objects, parts without requiring segmentation; robustness to clutter & occlusion
- Robustness: similar descriptors in spite of moderate view changes, noise, blur, etc.

Once we have the features themselves, how to use for recognition, search?

Visual Object Recognition Tutorial

Basic flow



Index each one into pool of descriptors from previously seen images

Detect or sample features

Describe features

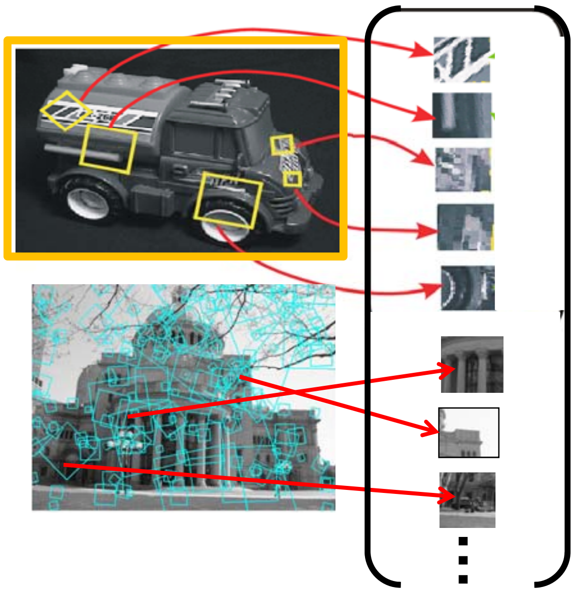
List of positions, scales, orientations → Associated list of d-dimensional descriptors

Pose clustering and verification with SIFT

To detect **instances** of objects from a model base:

1) Index descriptors (distinctive features narrow possible matches)

Indexing local features

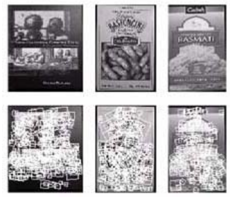


Pose clustering and verification with SIFT

To detect **instances** of objects from a model base:

1) Index descriptors (distinctive features narrow possible matches)
2) Generalized Hough transform to vote for poses (keypoints have record of parameters relative to model coordinate system) *[next week]*
3) Affine fit to check for agreement between model and image features (approximates perspective projection for planar objects)

Planar objects



Model images and their SIFT keypoints

Input image

Model keypoints that were used to recognize, get least squares solution.

Recognition result

[Lowe]

Visual Object Recognition Tutorial

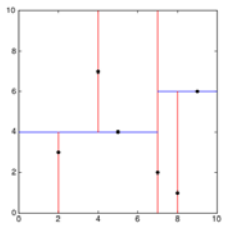
Indexing local features

- **With potentially thousands of features per image, and hundreds to millions of images to search, how to efficiently find those that are relevant to a new image?**
  - **Low-dimensional descriptors : can use standard efficient data structures for nearest neighbor search**
  - **High-dimensional descriptors: approximate nearest neighbor search methods more practical**
  - **Inverted file indexing schemes**

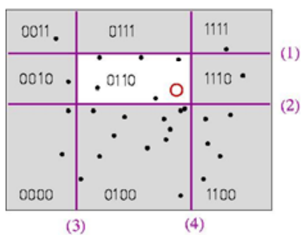
K. Grauman, B. Leibe

Visual Object Recognition Tutorial

Indexing local features: approximate nearest neighbor search



Best-Bin First (BBF), a variant of k-d trees that uses priority queue to examine most promising branches first [Beis & Lowe, CVPR 1997]



Locality-Sensitive Hashing (LSH), a randomized hashing technique using hash functions that map similar points to the same bin, with high probability [Indyk & Motwani, 1998]

41

Visual Object Recognition Tutorial

Indexing local features

- With potentially thousands of features per image, and hundreds to millions of images to search, how to efficiently find those that are relevant to a new image?
  - Low-dimensional descriptors : can use standard efficient data structures for nearest neighbor search
  - High-dimensional descriptors: approximate nearest neighbor search methods more practical
  - Inverted file indexing schemes

K. Grauman, B. Leibe

Visual Object Recognition Tutorial

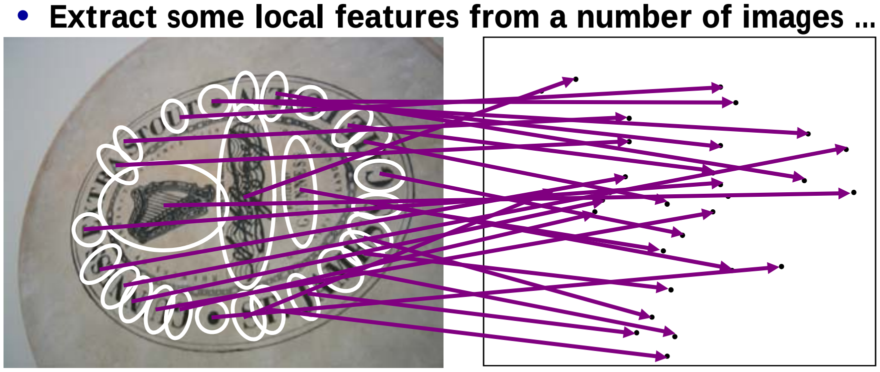
Indexing local features: inverted file index

- For text documents, an efficient way to find all *pages* on which a *word* occurs is to use an index…
- We want to find all *images* in which a *feature* occurs.
- To use this idea, we'll need to map our features to "visual words".

Visual Object Recognition Tutorial

Visual words: main idea

- Extract some local features from a number of images …



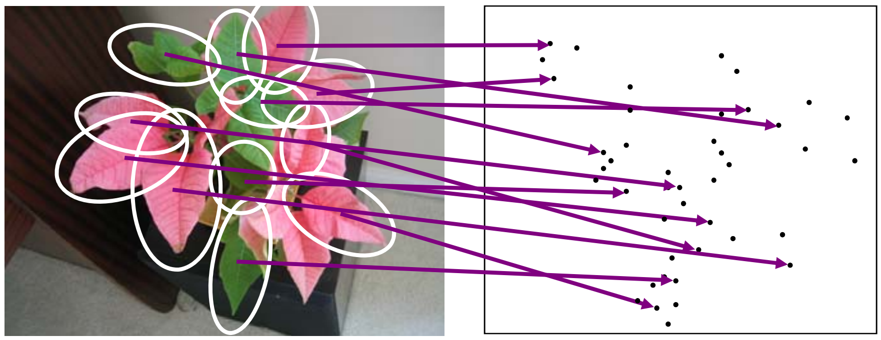
e.g., SIFT descriptor space: each point is 128-dimensional

Slide credit: D. Nister

44

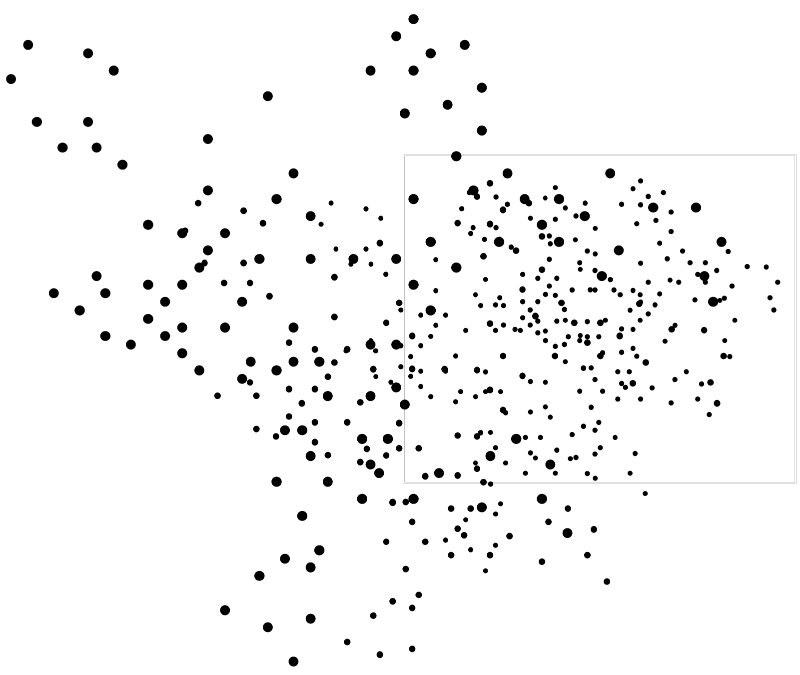
Visual Object Recognition Tutorial

Visual words: main idea

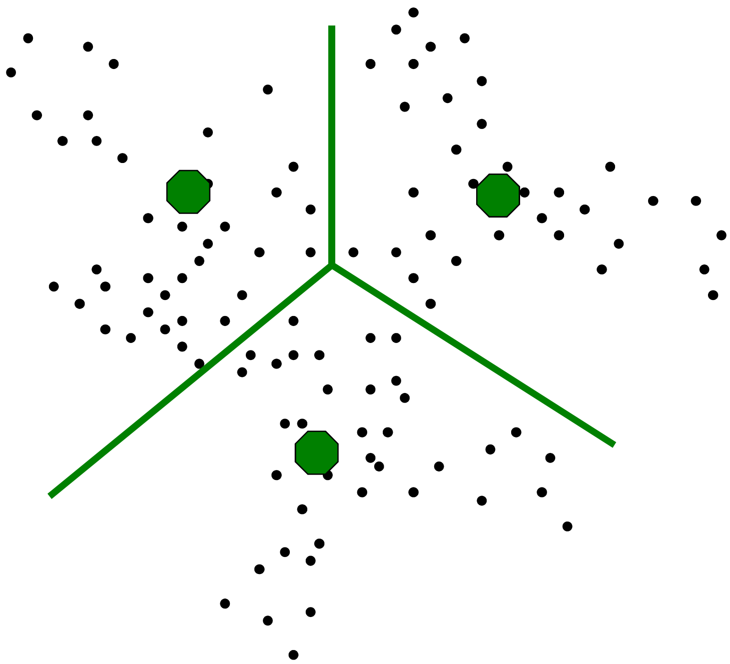


Slide credit: D. Nister

Visual Object Recognition Tutorial



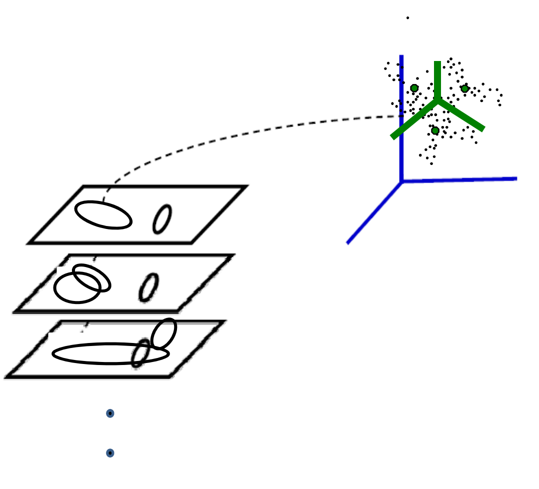
Slide credit: D. Nister



Slide credit: D. Nister

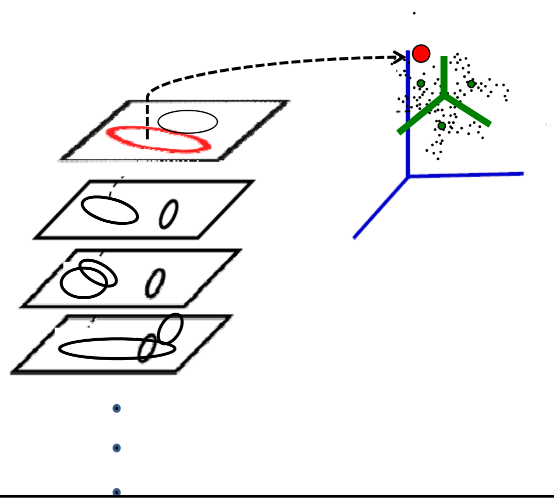
Visual words: main idea

Map high-dimensional descriptors to tokens/words by quantizing the feature space



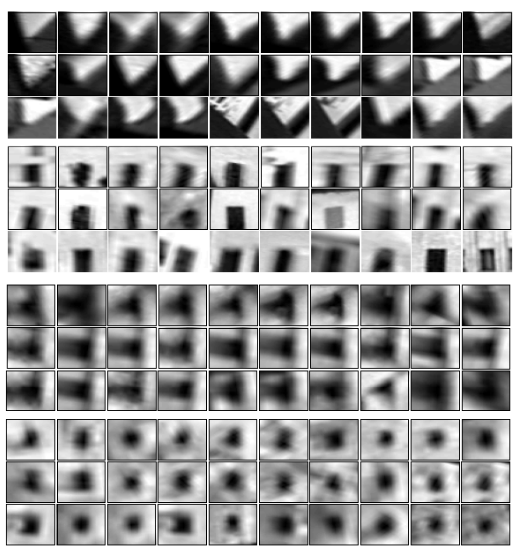
Visual words: main idea

Map high-dimensional descriptors to tokens/words by quantizing the feature space



Visual words

- **Example: each group of patches belongs to the same visual word**



Visual Object Recognition Tutorial

Inverted file index for images comprised of visual words

frame #5

frame #10

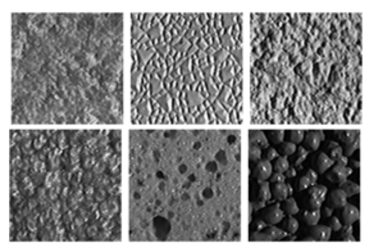
Word number	List of image numbers
1	5,10, ...
2	10,...
...	...

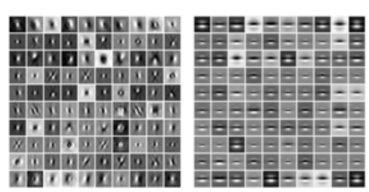
Image credit: A. Zisserman

Visual words

- **First explored for texture and material representations**
- ***Texton* = cluster center of filter responses over collection of images**
- **Describe textures and materials based on distribution of prototypical texture elements.**

Leung & Malik 1999; Varma & Zisserman, 2002; Lazebnik, Schmid & Ponce, 2003;

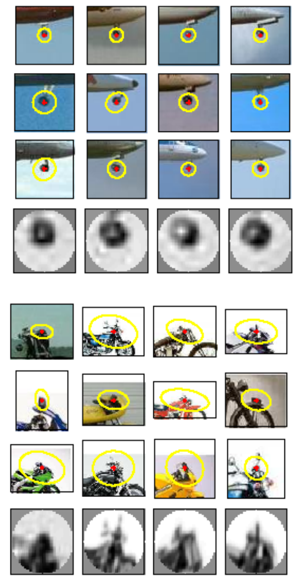




Visual words

- **More recently used for describing scenes and objects for the sake of indexing or classification.**

Sivic & Zisserman 2003; Csurka, Bray, Dance, & Fan 2004; many others.



Visual Object Recognition Tutorial

53

Visual vocabulary formation

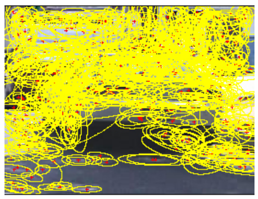
Issues:

- **Sampling strategy: where to extract features?**
- **Clustering / quantization algorithm**
- **Unsupervised vs. supervised**
- **What corpus provides features (universal vocabulary?)**
- **Vocabulary size, number of words**

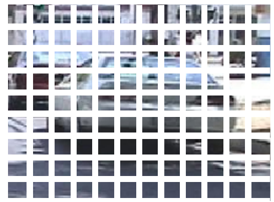
Visual Object Recognition Tutorial

Visual Object Recognition Tutorial

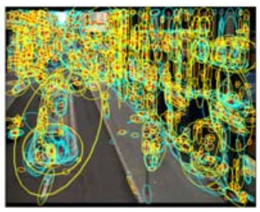
Sampling strategies



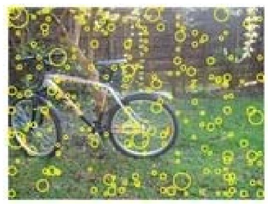
Sparse, at interest points



Dense, uniformly



Multiple interest operators



Randomly

Image credits: F-F. Li, E. Nowak, J. Sivic

- To find specific, textured objects, sparse sampling from interest points often more reliable.
- Multiple complementary interest operators offer more image coverage.
- For object categorization, dense sampling offers better coverage.

See [Nowak, Jurie & Triggs, ECCV 2006] **, and Gautam's demo!**

Visual Object Recognition Tutorial

Clustering / quantization methods

- **k-means (typical choice), agglomerative clustering, mean-shift,…**
- **Hierarchical clustering: allows faster insertion / word assignment while still allowing large vocabularies**
  - **Vocabulary tree [Nister & Stewenius, CVPR 2006]**

Visual Object Recognition Tutorial

Visual vocabulary formation

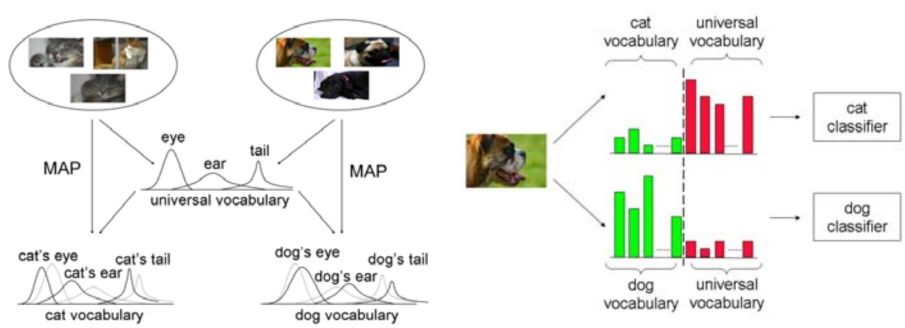
Issues:

- **Sampling strategy: where to extract features?**
- **Clustering / quantization algorithm**
- **Unsupervised vs. supervised**
- **What corpus provides features (universal vocabulary?)**
- **Vocabulary size, number of words**

Visual Object Recognition Tutorial

Supervised vocabulary formation

- **Recent work considers how to leverage labeled images when constructing the vocabulary**

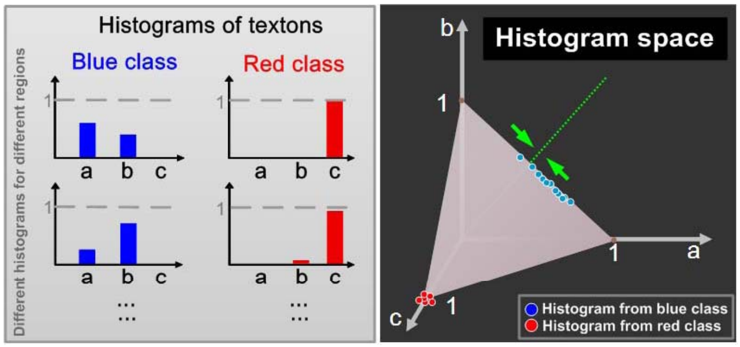


Perronnin, Dance, Csurka, & Bressan, Adapted Vocabularies for Generic Visual Categorization, ECCV 2006.

58

Supervised vocabulary formation

- **Merge words that don't aid in discriminability**

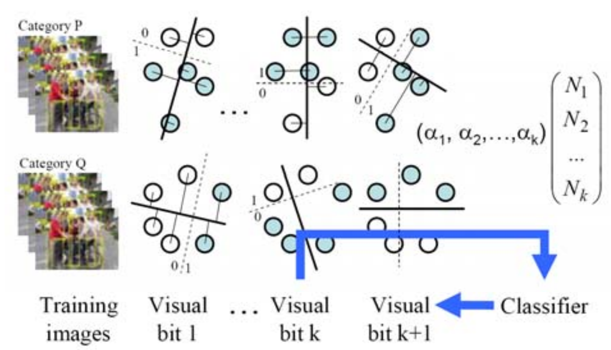


Winn, Criminisi, & Minka, Object Categorization by Learned Universal Visual Dictionary, ICCV 2005

Visual Object Recognition Tutorial

Supervised vocabulary formation

- **Consider vocabulary and classifier construction jointly.**

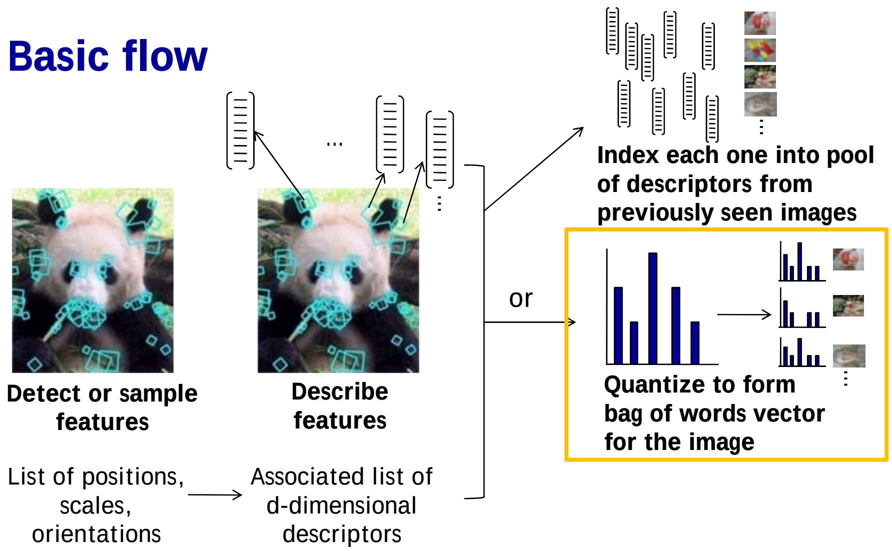


Yang, Jin, Sukthankar, & Jurie, Discriminative Visual Codebook Generation with Classifier Training for Object Category Recognition, CVPR 2008.

60

Visual Object Recognition Tutorial

Basic flow



Visual Object Recognition Tutorial

Bags of visual words

- Summarize entire image based on its distribution (histogram) of word occurrences.
- Analogous to bag of words representation commonly used for documents.
- Set of patches -> vector
- Good empirical results for image classification.

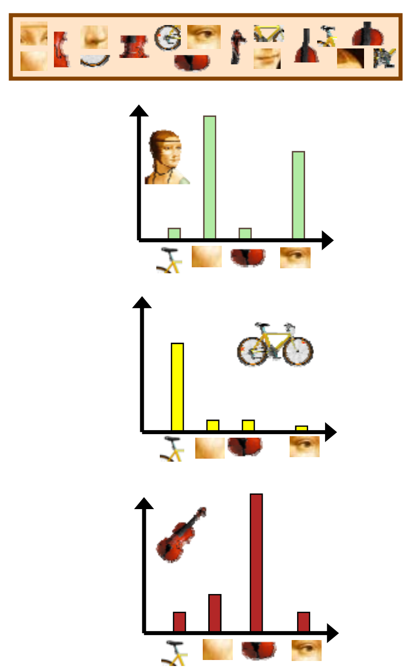


Image credit: Fei-Fei Li

Bags of visual words for image recognition

Caltech 6 dataset

class	bag of features Zhang et al. (2005)	bag of features Willamowski et al. (2004)	Parts-and-shape model Fergus et al. (2003)
airplanes	**98.8**	97.1	90.2
cars (rear)	98.3	**98.6**	90.3
cars (side)	**95.0**	87.3	88.5
faces	**100**	99.3	96.4
motorbikes	**98.5**	98.0	92.5
spotted cats	**97.0**	—	90.0

Source: Lana Lazebnik

Visual Object Recognition Tutorial

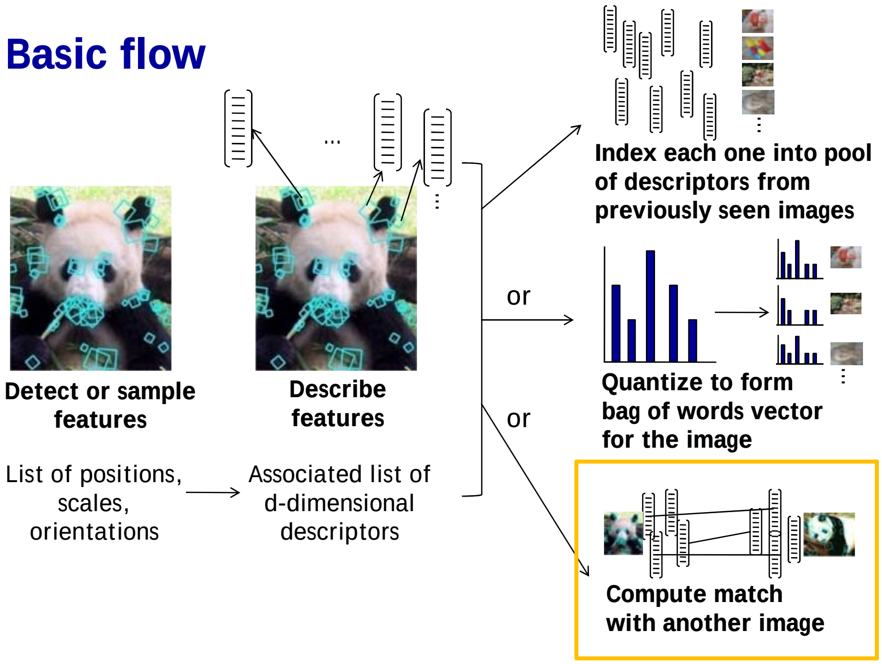
Bags of words: pros and cons

+ flexible to geometry / deformations / viewpoint
+ compact summary of image content
+ provides vector representation for sets
+ has yielded good recognition results in practice

- basic model ignores geometry – must verify afterwards, or encode via features
- background and foreground mixed when bag covers whole image
- interest points or sampling: no guarantee to capture object-level parts
- optimal vocabulary formation remains unclear

Visual Object Recognition Tutorial

Basic flow



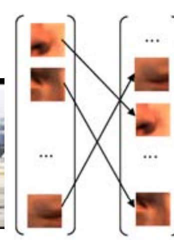
Visual Object Recognition Tutorial

Local feature correspondences

- The matching between sets of local features helps to establish overall similarity between objects or shapes.
- Assigned matches also useful for localization

Shape context [Belongie & Malik 2001]

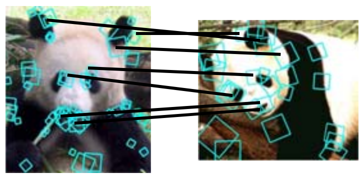
Low-distortion matching [Berg & Malik 2005]



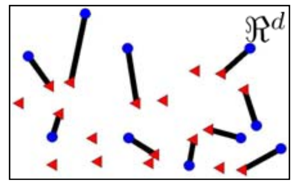
Match kernel [Wallraven, Caputo & Graf 2003]

Local feature correspondences

- **Least cost match: minimize total cost between matched points**



$\mathbf{X} = \{\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_m\}$ $\mathbf{Y} = \{\vec{\mathbf{y}}_1, \ldots, \vec{\mathbf{y}}_n\}$



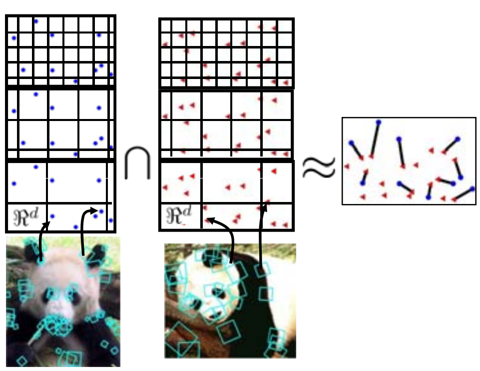
$$\min_{\pi: X \to Y} \sum_{x_i \in X} \left\| x_i - \pi(x_i) \right\|$$

- **Least cost *partial* match: match all of smaller set to some portion of larger set.**

Visual Object Recognition Tutorial

Pyramid match kernel (PMK)

- **Optimal matching expensive relative to number of features per image (m).**
- **PMK is approximate partial match for efficient discriminative learning from <u>sets</u> of local features.**



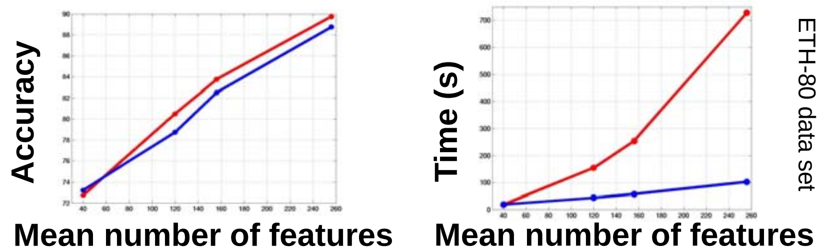
Optimal match: $O(m^3)$
Greedy match: $O(m^2 \log m)$
Pyramid match: $O(m)$

[Grauman & Darrell, ICCV 2005]

Visual Object Recognition Tutorial

Pyramid match kernel

- **Forms a Mercer kernel -> allows classification with SVMs, use of other kernel methods**
- **Bounded error relative to optimal partial match**
- **Linear time -> efficient learning with large feature sets**

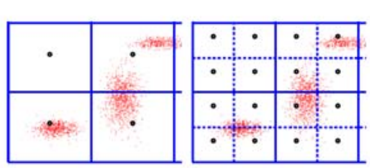


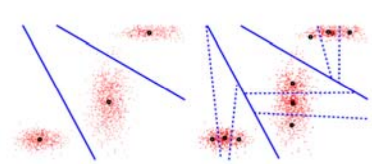
Match *[Wallraven et al.]*	$O(m^2)$
Pyramid match	$O(m)$

Visual Object Recognition Tutorial

Pyramid match kernel

- **Forms a Mercer kernel -> allows classification with SVMs, use of other kernel methods**
- **Bounded error relative to optimal partial match**
- **Linear time -> efficient learning with large feature sets**
- **Use data-dependent pyramid partitions for high-d feature spaces**

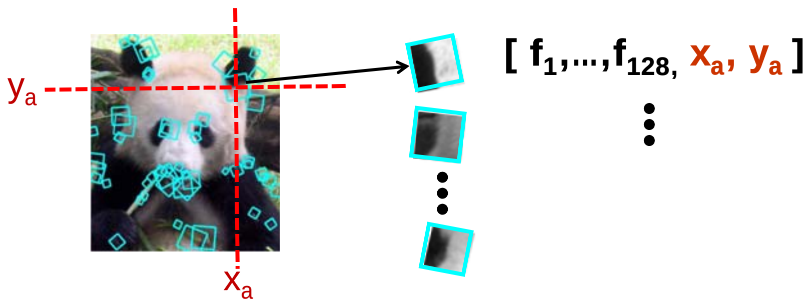

Uniform pyramid bins


Vocabulary-guided pyramid bins

Code for PMK: http://people.csail.mit.edu/jjl/libpmk/

Matching smoothness & local geometry

- Solving for linear assignment means (non-overlapping) features can be matched independently, ignoring relative geometry (as in bag of words model).
- One alternative: simply expand feature vectors to include spatial information *before* matching.



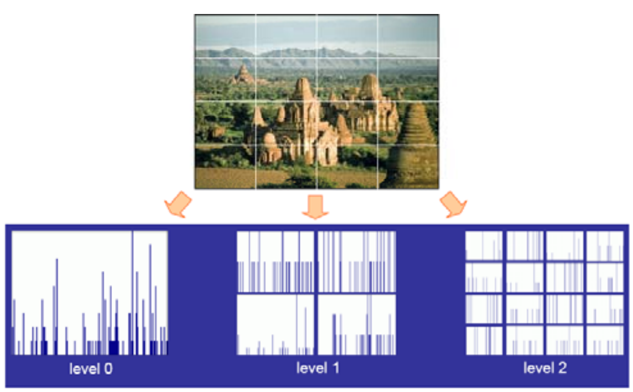
Visual Object Recognition Tutorial

K. Grauman, B. Leibe

71

Spatial pyramid match kernel

- First quantize descriptors into words, then do one pyramid match *per word* in image coordinate space.



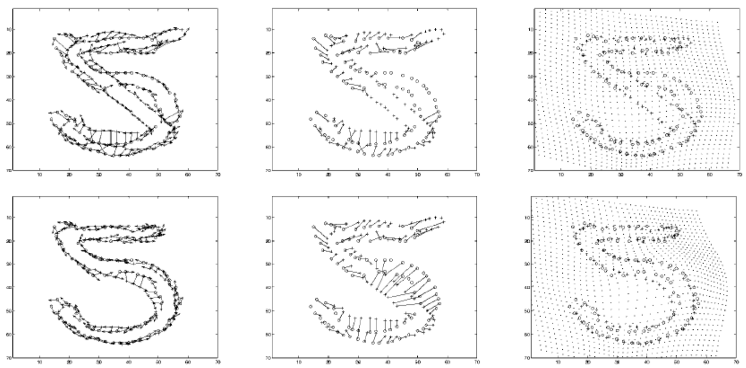
Lazebnik, Schmid & Ponce, CVPR 2006

Visual Object Recognition Tutorial

K. Grauman, B. Leibe

Matching smoothness & local geometry

- **Use correspondence to estimate parameterized transformation, regularize to enforce smoothness**



Shape context matching [Belongie, Malik, & Puzicha 2001]

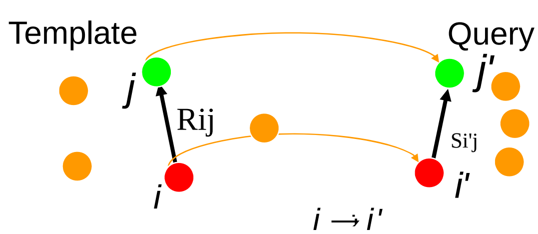
Code: http://www.eecs.berkeley.edu/Research/Projects/CS/vision/shape/sc_digits.html

K. Grauman, B. Leibe

Visual Object Recognition Tutorial

Matching smoothness & local geometry

- **Let matching cost include term to penalize distortion between pairs of matched features.**



Approximate for efficient solutions: Berg & Malik, CVPR 2005; Leordeanu & Hebert, ICCV 2005

Figure credit: Alex Berg

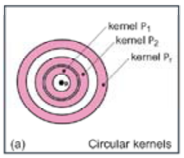
K. Grauman, B. Leibe

Visual Object Recognition Tutorial

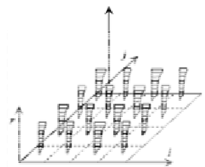
Visual Object Recognition Tutorial

Matching smoothness & local geometry

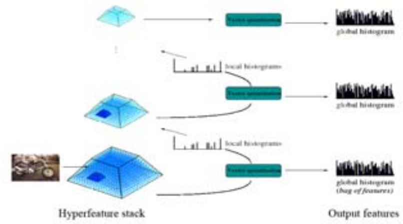
- **Compare "semi-local" features: consider configurations or neighborhoods and co-occurrence relationships**



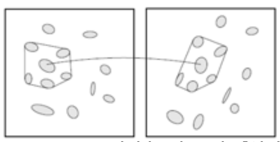
Correlograms of visual words [Savarese, Winn, & Criminisi, CVPR 2006]



Proximity distribution kernel [Ling & Soatto, ICCV 2007]



Hyperfeatures: Agarwal & Triggs, ECCV 2006]



Feature neighborhoods [Sivic & Zisserman, CVPR 2004]

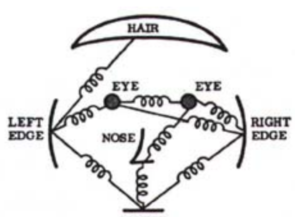
Tiled neighborhood [Quack, Ferrari, Leibe, van Gool ICCV 2007]

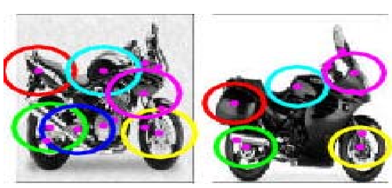
K. Grauman, B. Leibe

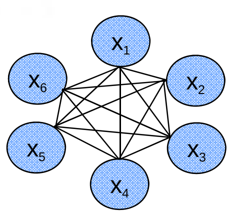
Visual Object Recognition Tutorial

Matching smoothness & local geometry

- **Learn or provide explicit object-specific shape model [Next week: part-based models]**







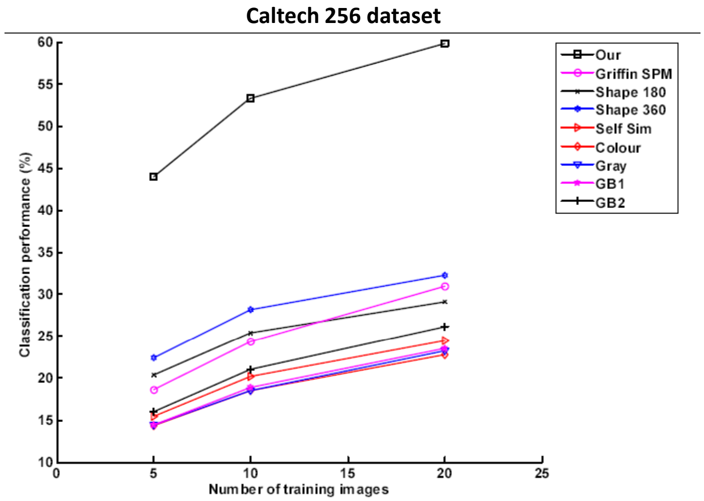
Distance/metric/kernel learning

- Exploit partially labeled data and/or (dis)similarity constraints to construct more useful distance function
- Number of existing techniques

Distance/metric/kernel learning

- "Multiple kernel learning": Optimize weights on kernel matrices, where each matrix is from a different feature type or similarity measure. "Align" to the optimal kernel matrix.
  - [e.g. Varma & Ray ICCV 2007, Bosch et al. CIVR 2007, Kumar & Sminchisescu ICCV 2007]
- Example-based distance learning: Optimize weights on each feature within a training image
  - [Frome et al. ICCV 2007]
- Learn metric based on similarity / in-class constraints
  - Often Mahalanobis distances [e.g. Hertz et al. CVPR 2004, Kumar et al. ICCV 2007, Jain et al. CVPR 2008]

Example: impact of kernel combination



Varma and Ray, ICCV 2007

Example Applications: Local Feature Matching

Sony Aibo
(Evolution Robotics)

SIFT usage

- Recognize docking station
- Communicate with visual cards

Other uses

- Place recognition
- Loop closure in SLAM

Slide credit: David Lowe

Example Applications: Local Feature Matching

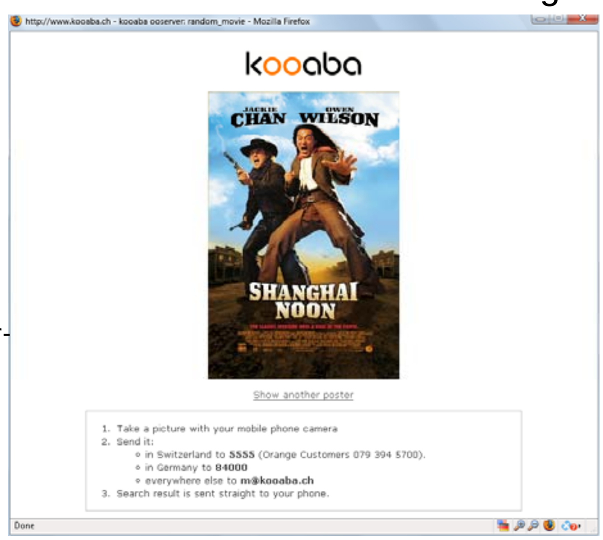
Mobile tourist guide
- **Self-localization**
- **Object/building recognition**
- **Photo/video augmentation**

[Quack, Leibe, Van Gool, CIVR'08]

Example Applications: Local Feature Matching

50'000 movie posters indexed

Query-by-image from mobile phone available in Switzer-land



http://www.kooaba.com/en/products_engine.html#

K. Grauman, B. Leibe

Summary

- **Local features are a useful, flexible representation**
  - **Invariance properties - typically built into the descriptor**
  - **Distinctive, especially helpful for identifying specific textured objects**
  - **Breaking image into regions/parts gives tolerance to occlusions and clutter**
  - **Mapping to visual words forms discrete tokens from image regions**

- **Efficient methods available for**
  - **Indexing patches or regions**
  - **Comparing distributions of visual words**
  - **Matching features**